Ex.2 Working with Data using Pandas

September 4, 2023

Expt: No 2

URK21CS1128

Bewin Felix R A

[ ]:
```
Aim: To execute the basic functionalities using pandas with data.
```

Description: Python Pandas is defined as an open-source library that provides high-performance data manipulation in Python. Started by Wes McKinney in 2008 out of a need for a powerful and flexible quantitative analysis tool, panda has grown into one of the most popular Python libraries. Pandas is built on top of the Numpy package, means Numpy is required for operating the Pandas.Pandas DataFrame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the data, rows, and columns. Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

Program:

[1]:
```
import pandas as pd
df = pd.read_csv("Titanic.csv")
df
```

[1]:
```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
```

```
2                                Heikkinen, Miss. Laina  female  26.0      0
3          Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                              Allen, Mr. William Henry    male  35.0      0
..                                                  …     …   …    …
886                               Montvila, Rev. Juozas    male  27.0      0
887                        Graham, Miss. Margaret Edith  female  19.0      0
888            Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                               Behr, Mr. Karl Howell    male  26.0      0
890                                 Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …              …       …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

Q1: Display the columns that have null and its count

[5]:
```
print("URK21CS1128")
df0=df.isnull().sum().sum()
print("Count:",df0)
```

```
URK21CS1128
Count: 866
```

[6]:
```
df
```

[6]:
```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3
```

```
                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0
..                                                ...     …   …    …
886                            Montvila, Rev. Juozas    male  27.0      0
887                     Graham, Miss. Margaret Edith  female  19.0      0
888         Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                            Behr, Mr. Karl Howell    male  26.0      0
890                              Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …               …      …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

Q2: Display the statistical description of the numerical and non-numerical columns

[7]:
```
print("URK21CS1128")
df.describe()
```

```
URK21CS1128
```

[7]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
```

```
mean      0.381594   32.204208
std       0.806057   49.693429
min       0.000000    0.000000
25%       0.000000    7.910400
50%       0.000000   14.454200
75%       0.000000   31.000000
max       6.000000  512.329200
```

[8]:
```
print("URK21CS1128")
df.describe(include='object')
```

```
URK21CS1128
```

[8]:
```
                           Name   Sex  Ticket    Cabin Embarked
count                       891   891     891      204      889
unique                      891     2     681      147        3
top     Braund, Mr. Owen Harris  male  347082  B96 B98        S
freq                          1   577       7        4      644
```

Q3: Display the rows that 'Sex' column value is male and observe the count

[9]:
```
print("URK21CS1128")
df2 = df[df['Sex']=='male']
df2
```

```
URK21CS1128
```

[9]:
```
     PassengerId  Survived  Pclass                                Name   Sex  \
0              1         0       3             Braund, Mr. Owen Harris  male
4              5         0       3            Allen, Mr. William Henry  male
5              6         0       3                    Moran, Mr. James  male
6              7         0       1             McCarthy, Mr. Timothy J  male
7              8         0       3      Palsson, Master. Gosta Leonard  male
..           …       …     …                                     …   …
883          884         0       2       Banfield, Mr. Frederick James  male
884          885         0       3              Sutehall, Mr. Henry Jr  male
886          887         0       2               Montvila, Rev. Juozas  male
889          890         1       1                Behr, Mr. Karl Howell  male
890          891         0       3                  Dooley, Mr. Patrick  male

      Age  SibSp  Parch            Ticket     Fare Cabin Embarked
0    22.0      1      0         A/5 21171   7.2500   NaN        S
4    35.0      0      0            373450   8.0500   NaN        S
5     NaN      0      0            330877   8.4583   NaN        Q
6    54.0      0      0             17463  51.8625   E46        S
7     2.0      3      1            349909  21.0750   NaN        S
..    …    …     …               …       …   …       …
883  28.0      0      0  C.A./SOTON 34068  10.5000   NaN        S
884  25.0      0      0   SOTON/OQ 392076   7.0500   NaN        S
```

```
886  27.0      0      0            211536  13.0000   NaN        S
889  26.0      0      0            111369  30.0000  C148        C
890  32.0      0      0            370376   7.7500   NaN        Q

[577 rows x 12 columns]
```

[10]:
```
#URK21CS1128
df2.shape[0]
```

[10]:
```
577
```

Q4: Display the Name and Age of first 25 rows with 'Embarked' column is C (Cherbourg)

[11]:
```
print("URK21CS1128")
df4 = df[df['Embarked']=="C"]
df4 = df4.head(25)[['Name','Age']]
df4
```

```
URK21CS1128
```

[11]:

	Name	Age
1	Cumings, Mrs. John Bradley (Florence Briggs Th…	38.0
9	Nasser, Mrs. Nicholas (Adele Achem)	14.0
19	Masselmani, Mrs. Fatima	NaN
26	Emir, Mr. Farred Chehab	NaN
30	Uruchurtu, Don. Manuel E	40.0
31	Spencer, Mrs. William Augustus (Marie Eugenie)	NaN
34	Meyer, Mr. Edgar Joseph	28.0
36	Mamee, Mr. Hanna	NaN
39	Nicola-Yarred, Miss. Jamila	14.0
42	Kraeff, Mr. Theodor	NaN
43	Laroche, Miss. Simonne Marie Anne Andree	3.0
48	Samaan, Mr. Youssef	NaN
52	Harper, Mrs. Henry Sleeper (Myna Haxtun)	49.0
54	Ostby, Mr. Engelhart Cornelius	65.0
57	Novel, Mr. Mansouer	28.5
60	Sirayanian, Mr. Orsen	22.0
64	Stewart, Mr. Albert A	NaN
65	Moubarek, Master. Gerios	NaN
73	Chronopoulos, Mr. Apostolos	26.0
96	Goldschmidt, Mr. George B	71.0
97	Greenfield, Mr. William Bertram	23.0
111	Zabour, Miss. Hileni	14.5
114	Attalah, Miss. Malake	17.0
118	Baxter, Mr. Quigg Edmond	24.0
122	Nasser, Mr. Nicholas	32.5

Q5: Display the rows that Age>20 and Survived status is 0

[12]:
```
print("URK21CS1128")
df5=df[(df.Age>20)&(df.Survived==0)]
df5
```

```
URK21CS1128
```

[12]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande…	female	31.0	1	0	345763	18.0000	NaN	S
..	…	…	…	…	…	…	…	…	…	…	…	…
883	884	0	2	Banfield, Mr. Frederick James	male	28.0	0	0	C.A./SOTON 34068	10.5000	NaN	S
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.0500	NaN	S
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

```
[327 rows x 12 columns]
```

Q6: Display the top 10 rows of the ‘Age’ column with NAN value

[13]:
```
print("URK21CS1128")
df6 = df[df.Age.isna()]
d6 = df6.head(10)
d6
```

URK21CS1128

[13]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000	NaN	S
19	20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	2649	7.2250	NaN	C
26	27	0	3	Emir, Mr. Farred Chehab	male	NaN	0	0	2631	7.2250	NaN	C
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NaN	0	0	330959	7.8792	NaN	Q
29	30	0	3	Todoroff, Mr. Lalio	male	NaN	0	0	349216	7.8958	NaN	S
31	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NaN	1	0	PC 17569	146.5208	B78	C
32	33	1	3	Glynn, Miss. Mary Agatha	female	NaN	0	0	335677	7.7500	NaN	Q
36	37	1	3	Mamee, Mr. Hanna	male	NaN	0	0	2677	7.2292	NaN	C
42	43	0	3	Kraeff, Mr. Theodor	male	NaN	0	0	349253	7.8958	NaN	C

Q7: Display the max value in 'Fare', min value in 'Age', and mean value in 'Fare'

[14]:
```
print("URK21CS1128")
df7=df['Fare'].max()
print(df7)
```

```
d7 = df['Age'].min()
print(d7)
d= df['Fare'].mean()
print(d)
```

```
URK21CS1128
512.3292
0.42
32.204207968574636
```

Q8: Display unique values in the Embarked column

[15]:
```
print("URK21CS1128")
df8 = df['Embarked'].unique()
df8
```

```
URK21CS1128
```

[15]:
```
array(['S', 'C', 'Q', nan], dtype=object)
```

Q9: Update the data frame with new column 'New_Fare'. New_Fare = Fare + 100 and observe the size of the data frame

[16]:
```
df['New_Fare'] = df['Fare']+100
print(df)
df.shape[0]
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                                Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                             Allen, Mr. William Henry    male  35.0      0
..                                                 …     …   …    …
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
```

```
889                             Behr, Mr. Karl Howell    male  26.0      0
890                               Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked  New_Fare
0        0         A/5 21171   7.2500   NaN        S  107.2500
1        0          PC 17599  71.2833   C85        C  171.2833
2        0  STON/O2. 3101282   7.9250   NaN        S  107.9250
3        0            113803  53.1000  C123        S  153.1000
4        0            373450   8.0500   NaN        S  108.0500
..     …               …       …   …      …       …
886      0            211536  13.0000   NaN        S  113.0000
887      0            112053  30.0000   B42        S  130.0000
888      2        W./C. 6607  23.4500   NaN        S  123.4500
889      0            111369  30.0000  C148        C  130.0000
890      0            370376   7.7500   NaN        Q  107.7500

[891 rows x 13 columns]
```

[16]:
```
891
```

Q10: Drop the New_Fare column permanently and observe the size of the data frame

[17]:
```
df10 = df.drop('New_Fare',axis=1,inplace=True)
print(df10)
df.shape[0]
```

```
None
```

[17]:
```
891
```

Q11: Drop the rows with NAN and observe the size of the data frame

[18]:
```
df.dropna(inplace=True)
df
```

[18]:
```
     PassengerId  Survived  Pclass  \
1              2         1       1
3              4         1       1
6              7         0       1
10            11         1       3
11            12         1       1
..           …       …     …
871          872         1       1
872          873         0       1
879          880         1       1
887          888         1       1
889          890         1       1

                                                  Name     Sex   Age  SibSp  \
```

```
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
6                              McCarthy, Mr. Timothy J    male  54.0      0
10                     Sandstrom, Miss. Marguerite Rut  female   4.0      1
11                            Bonnell, Miss. Elizabeth  female  58.0      0
..                                                 …     …   …    …
871   Beckwith, Mrs. Richard Leonard (Sallie Monypeny)  female  47.0      1
872                            Carlsson, Mr. Frans Olof    male  33.0      0
879      Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)  female  56.0      0
887                        Graham, Miss. Margaret Edith  female  19.0      0
889                              Behr, Mr. Karl Howell    male  26.0      0

     Parch    Ticket     Fare        Cabin Embarked
1        0  PC 17599  71.2833          C85        C
3        0    113803  53.1000         C123        S
6        0     17463  51.8625          E46        S
10       1   PP 9549  16.7000           G6        S
11       0    113783  26.5500         C103        S
..     …       …       …           …      …
871      1     11751  52.5542          D35        S
872      0       695   5.0000  B51 B53 B55        S
879      1     11767  83.1583          C50        C
887      0    112053  30.0000          B42        S
889      0    111369  30.0000         C148        C

[183 rows x 12 columns]
```

Q12: Append two new rows in the data frame and observe the size of the data frame

[19]:
```
new_rows = [{'Name': 'Bewin', 'Age':20, 'Sex': 'male', 'Survived': 1,␣
 ↪'Embarked': 'S', 'Fare': 50},
            {'Name': 'John', 'Age': 25, 'Sex': 'male', 'Survived': 0,␣
 ↪'Embarked': 'C', 'Fare': 80}]
df = df.append(new_rows, ignore_index=True)
print(df.shape)
```

```
(185, 12)

/tmp/ipykernel_3879147/563540454.py:3: FutureWarning: The frame.append method is
deprecated and will be removed from pandas in a future version. Use
pandas.concat instead.
  df = df.append(new_rows, ignore_index=True)
```

Q13: Fill the NAN values in 'Cabin' column with 'A100' and observe the null values count

[20]:
```
print("URK21CS1128")
d13=df['Cabin'].fillna('A100', inplace=True)
df13 = df['Cabin'].isnull().sum()
print(df13)
```

URK21CS1128
0

Q14: Group the rows based on the 'Embarked' column and observe how many are C = Cherbourg, Q = Queenstown, S = Southampton

[21]:
```
grouped_embarked = df.groupby('Embarked').size()
print(grouped_embarked)
```

```
Embarked
C     66
Q      2
S    117
dtype: int64
```

Q15: Sort the data frame based on 'Fare'

[22]:
```
df.sort_values(by='Fare', inplace=True)
print(df)
```

```
     PassengerId  Survived  Pclass  \
169        807.0         0     1.0
46         264.0         0     1.0
179        873.0         0     1.0
148        716.0         0     3.0
12          76.0         0     3.0
..           …       …      …
86         439.0         0     1.0
13          89.0         1     1.0
7           28.0         0     1.0
137        680.0         1     1.0
153        738.0         1     1.0

                                          Name     Sex   Age  SibSp  Parch  \
169                      Andrews, Mr. Thomas Jr    male  39.0    0.0    0.0
46                        Harrison, Mr. William    male  40.0    0.0    0.0
179                     Carlsson, Mr. Frans Olof    male  33.0    0.0    0.0
148  Soholt, Mr. Peter Andreas Lauritz Andersen    male  19.0    0.0    0.0
12                       Moen, Mr. Sigurd Hansen    male  25.0    0.0    0.0
..                                           …     …   …    …    …
86                             Fortune, Mr. Mark    male  64.0    1.0    4.0
13                    Fortune, Miss. Mabel Helen  female  23.0    3.0    2.0
7                  Fortune, Mr. Charles Alexander    male  19.0    3.0    2.0
137           Cardeza, Mr. Thomas Drake Martinez    male  36.0    0.0    1.0
153                        Lesurer, Mr. Gustave J    male  35.0    0.0    0.0

       Ticket      Fare        Cabin Embarked
169    112050    0.0000          A36        S
46     112059    0.0000          B94        S
179       695    5.0000  B51 B53 B55        S
```

```
148    348124    7.6500        F G73        S
12     348123    7.6500        F G73        S
..        …       …          …      …
86      19950  263.0000  C23 C25 C27        S
13      19950  263.0000  C23 C25 C27        S
7       19950  263.0000  C23 C25 C27        S
137  PC 17755  512.3292  B51 B53 B55        C
153  PC 17755  512.3292         B101        C

[185 rows x 12 columns]
```

Result: The basic functionalities of python Data structures and data set using pandas were executed successfully.

[ ]:

[ ]: